PHIL 7001: Final Exam

Professor Boris Babic

Instructions:

This is the final examination for PHIL 7001. The assignment will be graded out of a total of 100 points. It is closed book, but you can have one double spaced cheat sheet, on which you can write anything you like.

You have 180 minutes to complete this test. If you finish early, you may submit your answers to the test invigilators and leave early. However, in order to avoid too much commotion, during the last 30 minutes please do not leave. That is, if you finish with less than 30 minutes remaining, please remain at your desk quietly until the exam is over.

Please make sure to write as clearly and legibly as possible. This will make it much easier for grading purposes.

You can write your answer in the exam booklet itself. Directly below each question. If you need more space, you can use the back side of each paper of the exam booklet.

Calculators are allowed, and you can use any calculator you wish. Smart devices of any kind are not allowed.

Please write your name and student number clearly on the first page of your submission.

There should be no trick questions on the exam. If anything seems unclear or ambiguous, then you can use your judgment to resolve the ambiguity, and note that you have done so in your answer. It is always best to state your assumptions as clearly as possible.

Problem 1: Logistic Regression (35 Points)

Suppose that you are working with a dataset called 'loan_data.csv', which contains information about loan applications. This dataset includes information such as an applicant's annual income (in dollars), credit score, and loan amount (in dollars). Your goal is to build a logistic regression classifier that predicts whether a loan application will be approved or denied. This is a binary classification task where '1' represents approval, and '0' represents denial. For the purpose of this analysis, we will assume a statistical significance level of 5%. Below you can see the first several rows of this dataset.

approval <int>	income <dbl>	credit_score <dbl>	loan_amount <dbl>
1	40214	656	43737
1	24933	664	41874
1	55531	672	35736
1	72686	671	27285
1	65865	741	29671
0	29733	632	45188

1A. Before fitting the classifier, your first task is to split the loan data into a training set and a testing set. Suppose we use 20% of the data as our training set, and 80% of the data as our test set. Is this a good idea? Explain in 1-2 sentences why or why not. (3 points)

1B. Suppose we've split the loan data into a training set 'loan_train', and a testing set 'loan_test'. You then need to fit a logistic regression model to predict loan approval, where income and credit score are the predictors. Please complete the blanks. (4 points)

```
logistic_model <- glm(approval ~ _ _ _ _ _ _ + _ _ _ _ _ _,
                      data = _ _ _ _ _ _,
                      family = '_ _ _ _ _ _')
```

1C. The following is the full summary of the fitted logistic regression model from above.

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.570e+01  1.132e+01  -6.687 2.28e-11 ***
## income       -1.013e-05  1.764e-05  -0.575    0.566
## credit_score  1.171e-01  1.746e-02   6.707 1.98e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 301.05  on 349  degrees of freedom
## Residual deviance:  93.45  on 347  degrees of freedom
## AIC: 99.45
```

Are income and credit score statistically significant predictors for the approval status of the loan application? Explain in 1-2 sentences why or why not, and circle above any quantities relevant to your answer. (4 points)

1D. To assess the classifier's performance, we will validate the model on the testing set. To do this, please complete the blanks below (2 points).

```
# Predict probabilities on the testing set
predicted_probs <- predict(_ _ _ _ _ _,
                           newdata = _ _ _ _ _ _,
                           type = "response")
```

1E. Consider the code below. Please explain in 2-3 sentences what this code accomplishes and why we would need it in order to use a logistic regression model to accomplish a classification task. (4 points)

```
predicted_labels <- ifelse(predicted_probs > 0.5, 1, 0)
```

1F. Now we are going to change the model a little bit. Suppose that the parameter coefficient estimate for the intercept is -7.57, the parameter coefficient estimate for income is -0.013. And suppose that the parameter coefficient estimate for credit score is 0.772. Using these three coefficient estimates, calculate the probability of a loan application being approved when the applicant's annual income is $35000 with a credit score of 600. (10 points).

1G. Suppose that you have also trained a Support Vector Machine (SVM) classifier for the loan dataset and you want to compare the performance of the two models. To assess their performance, you calculate the confusion matrix for both models and obtain the following.

Classifier	Accuracy	Sensitivity	Specificity	F1 score
Logistic	0.65	0.51	0.83	0.59
SVM	0.73	0.53	0.85	0.64

Based on the table above, which classifier appears to be performing better, and why do you think it is performing better? (3 points)

1H. Consider a different problem now. Suppose there is a horse race, and only two horses are racing, Sea Biscuit and Secrtariat. You are told that the odds of Sea Biscuit winning the race are 1:5. Using this information, please calculate the probability of Secretariat winning the face? (5 points).

Problem 2: Model Analysis and Assessment (35 Points)

Suppose that we have trained a certain classifier and obtained the following confusion matrix. Calculate the following quantities. Please show your calculations, and round your answer to 3 decimal places. Note: '1' represents the positive class.

	Predicted 0	Predicted 1
Actual 0	47	19
Actual 1	8	26

2A. What is the Type I (False Positive) error rate? (3 points)

2B. What is the Type II (False Negative) error rate? (3 points)

2C. Calculate the sensitivity. (3 points)

2D. Calculate the specificity. (3 points)

2E. Calculate the precision. (3 points)

2F. Explain in 2-3 sentences what an ROC curve represents, how it relates to the confusion matrix, and why we use it to evaluate a model's quality. (6 points)

2G. Explain in 1-2 sentences the advantage of using a random forest classification model instead of a single decision tree. (4 points).

2H. Explain in 1-2 sentences the advantage of using a support vector machine (SVM) model instead of a logistic regression model. (4 points).

2I. Explain in 2-3 sentences the difference between a soft margin SVM and a hard margin SVM. (6 points).

Problem 3: Neural Networks and Reinforcement Learning. (15 points)

3A. What is forward propagation in the context of neural networks? (2 points)

a). The process of adjusting weights based on error rates

b). The flow of information from the input to the output layer

c). The method of selecting the best activation function

d). The technique of dividing data into training and test sets

3B. In RL, what of the following best describes the sequence followed by an agent? (2 points)

a). Observes a state, takes an action, leads to state transition, receives a reward.

b). Receives a reward, takes an action, leads to state transition, observes a state.

c). Takes an action, observes a state, leads to state transition, receives a reward.

d). Observes a state, receives a reward, leads to state transition, takes an action.

3C. True or false: A neural network is a semi supervised algorithm. (2 points)

3D. True or False: A neural network with one hidden layer can approximate many non-linear functions. (2 points)

3E. True or False: A random process in which the probability of each state is independent of every other state is an example of a Markov Process. (2 points)

3F. Please explain in 2-4 sentences what is the Universal Approximation Theorem and why it is important in machine learning. (5 points)

Problem 4: Large Language Models. (15 Points)

Suppose we have a dictionary containing words related to a movie: ['Tom', 'said', 'the', 'movie', 'exciting', 'Mary', 'is', 'liked', 'and']. We want to train a Language Model (LLM) and we need to start with a vector embedding of some phrases.

4A. Using one-hot embedding, as we learned in class, please represent the phrase "Tom said the movie is exciting" in its vector notation. (4 points)

4B. Briefly explain in 1-2 sentences the purpose of a Multi-Layer Perceptron (MLP) in the context of LLMs. (3 points)

4C. Suppose we have reached the Unembedding step in the following (slightly shorter) dictionary. The final prediction matrix we generated is given as follows.

	Tom	said	the	movie	tonight	exciting	will
Tom	0.03	0.42	0.05	0.06	0.11	0.04	0.29
said	0.14	0.00	0.35	0.22	0.16	0.13	0.00
the	0.08	0.00	0.01	0.41	0.11	0.24	0.15
movie	0.03	0.11	0.01	0.01	0.58	0.08	0.18

Based on the table, what do you think the model would predict as the next token given the first two words "Tom said ... "? (4 points)

4D. Now, given the sentence "Tom said the movie ...", and using again the table from problem 4C, what will be the next token predicted, and what is the probability associated with this prediction? (4 points)